DOCUMENT RESUME

ED 219 428 TM 820 443

AUTHOR Nitko, Anthony J.
TITLE Properties of a Proposed Approximation to the Standard Error of Measurement.
PUB DATE Mar 82
NOTE 18p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (New York, NY, March 20-22, 1982).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Error of Measurement; *Estimation (Mathematics); *Mathematical Formulas; Statistical Analysis; Test Reliability
IDENTIFIERS *Garvin (A D)

ABSTRACT

An approximation formula for the standard error of measurement was recently proposed by Garvin. The properties of this approximation to the standard error of measurement are described in this paper and illustrated with hypothetical data. It is concluded that the approximation is a systematic overestimate of the standard error of measurement computed in the usual way with Kuder-Richardson formula 20. The relative error of the approximation was small for tests of more than 20 items. However, for short, internally consistent tests of the type used in instructional programs, the relative error can be quite large. (Author/BW)

Properties of a Proposed Approximation to the Standard Error of Measurement

Anthony J. Nitko
University of Pittsburgh

A Paper presented at the Annual Meeting of the National Council on Measurement in Education, New York City
March 19-23, 1982

Running Head: Properties of an Approximation

ABSTRACT

The properties of an approximation to the standard error of measurement were described and illustrated with hypothetical data. It was concluded that the approximation is a systematic overestimate of the standard error of measurement computed in the usual way with Kuder-Richardson formula 20. The relative error of the approximation was small for what was thought to represent many longer tests. However, for short, internally consistent tests of the type used in instructional programs, the relative error can be quite large.

Properties of a Proposed Approximation to the Standard Error of Measurement

The purpose of this paper is to examine some of the properties of an approximation formula for the standard error of measurement that was recently proposed by Garvin (1976). Examining the properties of this approximation would seem to be necessary because it has been recommended for use with classroom tests solely on the basis of its computational simplicity. Further, the empirical examples used to illustrate its use were not complete enough to judge the usefulness of the approximation for a wide range of classroom tests. Those using the proposed approximation may not be aware of its properties and recommendations for using it may well be tempered by a discussion of them.

The Proposed Approximation

The proposal is to approximate the standard error of measurement (SEM) by the following formula (Garvin, 1976, p. 102):

$$SEM = \frac{\sqrt{N\Sigma T - \Sigma T^2}}{N}, \qquad (1)$$

where N = the number of examinees taking the test

and T = the number of examinees answering a given item correctly.

The approximation is intended to apply to tests of k items, each of which is scored zero or one.

Formula (1) is derived by substituting N for N-1 and k for k-1 in the formula:

$$\text{SEM} = \hat{\sigma}\sqrt{1 - \text{KR20}}\ , \tag{2}$$

where

$$\hat{\sigma}^2 = \frac{\Sigma(X - \bar{X})^2}{N - 1}\ , \tag{3}$$

$$\text{KR20} = \frac{k}{k-1}\left(1 - \frac{\Sigma pq}{S^2}\right)\ , \tag{4}$$

and

$$S^2 = \frac{\Sigma(X - \bar{X})}{N}\ . \tag{5}$$

The symbols in these formulas have their usual meanings.

Some Properties of SEM

It should be noted that formula (2) is appropriate under certain conditions. One of these conditions is that KR20 is equal to the reliability of the test in question. The necessary and sufficient conditions under which this is true are called essential tau-equivalence (Novick & Lewis, 1967). If the true scores of the items of a test are not at least essentially tau-equivalent, KR20 will underestimate the test reliability, as defined in the classical sense, and the standard error of measurement will be overestimated. Additional problems exist: $\hat{\sigma}$ generally is not an unbiased estimate of the population standard deviation and KR20 is a biased estimate of its corresponding population value (Kristof, 1963). However, for many commercially available tests the standard error of measurement is determined using KR20. For classroom tests, most introductory testing and measurement texts express SEM in terms of S rather than $\hat{\sigma}$. This distinction will make a difference, as will be discussed below.

Although it is obvious, it should be stated that

$$SEM' = \sqrt{\Sigma pq} \quad , \tag{6}$$

where Σpq is the sum of the k item variances. If SEM′ is to be recommended, then an explanation of the relationship of the sum of the item variances to the total test error variance would seem to be in order.

Kuder-Richardson formula 20 in its general form is known as coefficient alpha (Cronbach, 1951). Using the notation of coefficient alpha and under the assumptions of at least essential tau-equivalence, it can be shown that

$$\Sigma\sigma_j^{\ 2} = \frac{\sigma_{T_x}^{\ 2}}{k} + \sigma_{E_x}^{\ 2} \quad , \tag{7}$$

where $\sigma_j^{\ 2}$ = the observed score variance item of j,

$\sigma_{T_x}^{\ 2}$ = the true score variance of the k-item test,

and $\sigma_{E_x}^{\ 2}$ = the error score variance of the k-item test.

While it is true that tests composed of items scored zero or one violate the assumptions under which equation (7) was derived (see, for example, Feldt, 1965), this expression would seem to hold well enough when Σpq is substituted for $\Sigma\sigma_j^{\ 2}$ to conclude that the square of SEM′ estimates something more than the error variance of the test. If expression (7) is true, then attempting to estimate the error variance via $(SEM')^2$ could be in serious error.

Classroom tests that would be used, say, to assess competency over small instructional units, would be relatively short and possibly quite internally consistent. Such short tests seem to be used quite frequently in the classroom. In such cases, the fraction $\sigma_{T_X}^2/k$ is likely to be high relative to $\sigma_{E_X}^2$. For example, Hsu (1971) reports data for four-item tests that measure attainment of single instructional objectives. Some of the KR20-values he reported were higher than .90. One test had KR20 equal to .97 (N = 49, S = 1.91). In this case the value of SEM′ is three times that of SEM (SEM′ = .997, SEM = .331).

To study how SEM′ differs systematically from SEM we need to express them in comparable terms. Manipulating formula (4) gives the following result:

$$(\mathrm{SEM}')^2 = S^2\left[1 - \left(\frac{k-1}{k}\right)\mathrm{KR20}\right] . \tag{8}$$

Garvin chose to express SEM in terms of $\hat{\sigma}$ instead of S. Since textbooks typically use S, both cases are examined below.

If SEM is expressed in terms of S, then it follows that

$$(\mathrm{SEM}')^2 - (\mathrm{SEM})^2 = \frac{S^2\mathrm{KR20}}{k} \tag{9}$$

$$\mathrm{SEM}' - \mathrm{SEM} = S\left[\sqrt{1 - \left(\frac{k-1}{k}\right)\mathrm{KR20}} - \sqrt{1-\mathrm{KR20}}\right] \tag{10}$$

and

$$\mathrm{SEM}' \geq \mathrm{SEM} \tag{11}$$

When the observed score variance of the test is computed as S for both KR20 and SEM, the approximation SEM′ is an overestimate of SEM except when KR20 = 0. For fixed test length k, the difference in the brackets of equation (10) is a monotonically increasing function of KR20. It increases rapidly at higher values of KR20 and gives a J-shaped appearance when graphed. When KR20 equals one, SEM′ is equal to $S/\sqrt{k}$, whereas, SEM equals zero.

If SEM is expressed in terms of $\hat{\sigma}$ and if KR20 is expressed in terms of S, then expression (10) becomes

$$SEM' - SEM = S\left[\sqrt{1 - \left(\frac{k-1}{k}\right)KR20} - \sqrt{\left(\frac{N}{N-1}\right)\left(1 - KR20\right)}\right] \quad (12)$$

In this case the bracketed difference is also a monotonically increasing, J-shaped function of KR20 for fixed test length k. However, the following relationships hold.

$$SEM' > SEM, \text{ when } \frac{k}{(N-1) + k} < KR20 \leq 1, \quad (13)$$

$$SEM' = SEM, \text{ when } KR20 = \frac{k}{(N-1) + k}, \quad (14)$$

and

$$SEM' < SEM, \text{ when } 0 \leq KR20 < \frac{k}{(N-1) + k}. \quad (15)$$

Alternately, we can write that

$$S\left(1 - \sqrt{\frac{N}{N-1}}\right) \leq (SEM' - SEM) \leq \frac{S}{\sqrt{k}} \quad (16)$$

when $0 \leq KR20 \leq 1$.

Relationship of SEM´ to Lord's Formulation

The values obtained for SEM´ in Garvin's article were contrasted to Lord's (1957) formulation of the standard error of measurement for individuals at a specific score point. Lord's formulation assumes that the k items of the test are a random sample from a very large domain of items. Under the conditions specified in Lord's development, the estimated error variance for individuals attaining a number right score of X_i is

$$\hat{\sigma}_{E_i}^2 = \frac{X_i (k - X_i)}{k - 1}. \qquad (17)$$

Since SEM´ is intended to approximate SEM, the value of comparing SEM´ to $\hat{\sigma}_{E_i}$ should be questioned. One way to interpret $(\mathrm{SEM})^2$ is as the average of all examinees' individual error score variances. If all individuals are measured with equal accuracy, then $(\mathrm{SEM})^2$ will apply equally well to each score-level; otherwise, it will not. Since $\hat{\sigma}_{E_i}$ reflects the idea that all persons are not measured equally well, it may be more useful to teachers than either SEM´ or SEM.

However, if one is to compare SEM´ with SEM, then to be consistent, one should compare SEM´ with an estimate based on the average of the $\hat{\sigma}_{E_i}$-values over all persons tested. Lord (1955) has shown that this average is

$$\mathrm{SEM}_L = S\sqrt{1 - \mathrm{KR21}}\,, \qquad (18)$$

where SEM_L = the estimated average standard error of measurement based on Lord's formulation,

and
$$KR21 = \left(\frac{k}{k-1}\right)\left(1 - \frac{\bar{X}(k - \bar{X})}{kS^2}\right) \quad (19)$$

The comparisons that are of interest are

$$(SEM')^2 - (SEM_L)^2 = S^2 \left[KR21 - \left(\frac{k-1}{k}\right) KR20\right] \quad (20)$$

and
$$(SEM_L)^2 - (SEM)^2 = S^2 (KR20 - KR21). \quad (21)$$

If all of the test items have the same difficulty value, then KR20 is equal to KR21 and

$$(SEM')^2 = \frac{\bar{X}(k - \bar{X})}{k}. \quad (22)$$

Under these special conditions SEM_L is identical to SEM; otherwise, SEM_L will be larger than SEM. The value of SEM′, however, will still maintain the relationships to SEM that are described by the equations in the preceding section.

Tucker (1949) has shown that, in general, KR20 is larger than KR21 by an amount equal to

$$\frac{k^2 S_p^2}{(k-1)S^2} \quad (23)$$

where S_p^2 is the variance of the item difficulties of the test. This means that the difference expressed in equation (20) is a function of the item difficulties of the test. We can express this difference as

$$(SEM')^2 - (SEM_L)^2 = \frac{S^2 KR20}{k} - \frac{k^2 S_p^2}{(k-1)} . \quad (24)$$

Similarly, we can rewrite equation (21) as

$$(SEM_L)^2 - (SEM)^2 = \frac{k^2 S_p^2}{(k-1)} . \quad (25)$$

By applying Tucker's (1949, formula 26) result along with equations (6) and (18), it can be shown that for k greater than one,

$$SEM' < SEM_L, \text{ when } \frac{KR20}{KR21} > \frac{k}{k-1} ; \quad (26)$$

$$SEM' = SEM_L, \text{ when } \frac{KR20}{KR21} = \frac{k}{k-1} ; \quad (27)$$

and

$$SEM' > SEM_L, \text{ when } \frac{KR20}{KR21} < \frac{k}{k-1} . \quad (28)$$

Taking into account equations (11) and (25) through (28), we can state the following relationships among the three estimators of the standard error of measurement:

$SEM_L > SEM' > SEM$, if condition (26) holds and if $KR20 > KR21$; (29)

$SEM_L = SEM' > SEM$, if condition (27) holds and if $KR20 > KR21$; (30)

and $SEM' > SEM_L > SEM$, if condition (28) holds and if $KR20 > KR21$. (31)

All three expressions are equal to S when $KR20 = KR21 = 0$. When $KR20 = KR21 \neq 0$, then SEM_L is equal to SEM, but SEM′ is still greater than SEM as shown by equation (10).

Representative Values of the Indices

Saupe (1961) has provided some representative values of test statistics for three general types of tests. Table 1 is based on Saupe's values and serves to illustrate the algebraic results obtained above. It should be noted that in making the calculations for Table 1, the values for KR20 and KR21 were carried to more decimal places than Saupe presented. Also, Table 1 uses expression (5) for the test variance for all computations.

Insert Table 1 about here

Two points may be noted from this table. First, as the average item difficulty level approaches .50 and as the variance of the item difficulties approaches zero, the discrepancies between all of the indices become smaller. Secondly, SEM_L tends to be closer to SEM than SEM′ is, when the variance of the item difficulties is less than .02, regardless of test length.

One would guess that most achievement tests would have distributions of item difficulties with values ranging between .20 and .80. A uniform distribution of item difficulties over this range would have $S_p^2 = .03$. A symetric, somewhat platykurtic distribution over the range .25 to .75, might be more typical of achievement tests designed to survey broad ranges of achievement in a subject. Such a distribution would likely have S_p^2 equal to about .01. If one were to concentrate item difficulties over a narrow range, say, .45 to .60, then a uniform distribution over this range would have S_p^2 less than .01. It is in the latter two cases that S_p^2 is smallest and SEM_L is closer to the value of SEM.

It should be noted, however, that the relative error of SEM′ is generally small for the values shown in Table 1, ranging from 8.8% (Δ_1 = .280) to 2.2% (Δ_1 = .048). The relative error for SEM_L is generally more substantial for these values, and ranges from 38.7% (Δ_2 = .555) to 0%. If it is true that most educational achievement tests would have $S_p^2 \leq .01$, then Table 1 would indicate that the relative error of SEM′ is small, being between 2% and 5%, when the test length is 20 items or more. The relative error for SEM_L is also small for these values of S_p^2 and test length, ranging between 0% and 3%.

Summary

Recently, SEM′ [as defined by formula (1)] was proposed as a computationally simple approximation to the standard error of measurement (SEM) for a test when this index is defined as in formula (2). Several properties of SEM′ were identified:

1. The index SEM′ can be shown to be systematically related to the true score variance of the test [formula (7)]. This means that for short, very reliable tests, the relative error in SEM′ can be quite high.

2. For the same data, SEM′ is always larger than SEM when KR20 > 0 and when the test's standard deviation is computed in the same way for both SEM and for KR20. When SEM is defined as in formula (2) and when KR20 is defined as in formula (4), then SEM′ can underestimate SEM for the same data.

3. It is felt that the comparison of SEM′ to Lord's $\hat{\sigma}_{E_i}$ was inappropriate, since SEM′ attempts to approximate the average examinee's error-score standard deviation, while $\hat{\sigma}_{E_i}$ does not. The "appropriate" comparison would be to SEM_L as defined in formula (18).

4. The relationship between SEM, SEM′, and SEM_L depends on the variance of the item difficult indices, or, alternatively, on the ratio of KR20 to KR21. These relationships are described by inequalities (26) through (31).

5. If it is true that most educational achievement tests have $S_p^2 \leq .01$, then the relative errors of both SEM′ and SEM_L in approximating SEM seem to be quite small when the number of items is over 20. The relative error of SEM_L is somewhat smaller than the relative error of SEM′ for this range of S_p^2-values, however.

Whether the information above argues for or against recommending the use of SEM′ for classroom tests depends on whether one is inclined to recommend computationally easier formulas that are known to be systematically biased and that seem to lack conceptual relationships to the qualities of the tests which they seek to estimate. If so, then SEM′ has merit, at least for longer tests with equal item difficulties.

References

Cronbach, L. J. Coefficient alpha and the internal structure of tests. Psychometrika, 1951, 16, 297-334.

Feldt, L. S. The approximate sampling distribution of Kuder-Richardson reliability coefficient twenty. Psychometrika, 1965, 30, 357-370.

Garvin, A. D. A simple, accurate approximation of the standard error of measurement. Journal of Educational Measurement, 1976, 13, 101-105.

Hsu, T. C. Empirical data on criterion-referenced tests. In M. R. Quilling (Chair), Criterion-referenced tests: Sense and nonsense. Symposium presented at the meeting of the American Educational Research Association, New York City, February, 1971.

Kristof, W. The statistical theory of stepped-up reliability coefficients when a test has been divided into several equivalent parts. Psychometrika, 1963, 28, 221-238.

Lord, F. M. Estimating test reliability. Educational and Psychological Measurement, 1955, 15, 325-336.

Lord, F. M. Do tests of the same length have the same standard errors of measurement? Educational and Psychological Measurement, 1957, 17, 510-521.

Novick, M. R. & Lewis, C. Coefficient alpha and the reliability of composite measurements. Psychometrika, 1967, 32, 1-13.

Saupe, J. L. Some useful estimates of the Kuder-Richardson formula number 20 reliability coefficient. Educational and Psychological Measurement, 1961, 21, 63-71.

Tucker, L. R. A note on the estimation of test reliability by the Kuder-Richardson formula (20). Psychometrika, 1949, 14, 117-119.

Table 1

Representative Values of SEH, SEH′, and SEH_L for 3 Types of Tests Described by Length and Variance

S_p^2	$\bar{p}$ or $\bar{q}$ = .3						$\bar{p}$ or $\bar{q}$ = .4						$\bar{p}$ or $\bar{q}$ = .5					
	KR20	SEH	SEH′	Δ_1	Δ_2	Δ_3	KR20	SEH	SEH′	Δ_1	Δ_2	Δ_3	KR20	SEH	SEH′	Δ_1	Δ_2	Δ_3
							TEST I. k = 20, S^2 = 9											
.09	.772	1.433	1.549	.116	.555	-.438	.702	1.638	1.732	.094	.501	-.408	.678	1.701	1.789	.087	.487	-.400
.04	.655	1.762	1.844	.082	.225	-.143	.585	1.933	2.000	.067	.207	-.140	.561	1.987	2.049	.063	.202	-.139
.02	.608	1.878	1.949	.072	.109	-.037	.538	2.039	2.098	.057	.101	-.042	.515	2.090	2.145	.055	.098	-.044
.01	.585	1.933	2.000	.067	.054	.013	.515	2.090	2.145	.055	.050	.005	.491	2.140	2.191	.051	.049	.002
.00	.561	1.987	2.049	.063	.000	.063	.491	2.140	2.191	.051	.000	.051	.468	2.188	2.236	.048	.000	.048
	KR21 = .561, SEH_L = 1.987						KR21 = .491, SEH_L = 2.140						KR21 = .468, SEH_L = 2.188					
							TEST II. k = 50, S^2 = 49											
.09	.895	2.263	2.449	.186	.853	-.667	.864	2.579	2.739	.159	.774	-.615	.854	2.676	2.828	.152	.752	-.600
.04	.843	2.770	2.915	.145	.347	-.201	.812	3.034	3.162	.128	.320	-.191	.802	3.117	3.240	.123	.312	-.188
.02	.823	2.949	3.082	.134	.168	-.035	.791	3.198	3.317	.119	.156	-.037	.781	3.276	3.391	.115	.152	-.037
.01	.812	3.034	3.162	.128	.083	.046	.781	3.276	3.391	.115	.077	.038	.771	3.353	3.464	.111	.072	.036
.00	.802	3.117	3.240	.124	.000	.124	.771	3.353	3.464	.111	.000	.111	.760	3.429	3.536	.107	.000	.107
	KR21 = .802, SEH_L = 3.117						KR21 = .771, SEH_L = 3.353						KR21 = .760, SEH_L = 3.429					
							TEST III. k = 100, S^2 = 196											
.09	.948	3.185	3.464	.280	1.201	-.921	.933	3.629	3.873	.244	1.089	-.845	.928	3.766	4.000	.234	1.058	-.824
.04	.922	3.898	4.123	.225	.488	-.267	.907	4.269	4.472	.203	.450	-.246	.902	4.385	4.583	.197	.439	-.242
.02	.912	4.149	4.359	.210	.237	-.027	.897	4.499	4.690	.191	.219	-.028	.892	4.610	4.796	.186	.214	-.028
.01	.907	4.269	4.472	.203	.117	.087	.892	4.610	4.796	.186	.108	.078	.886	4.718	4.899	.181	.106	.075
.00	.902	4.385	4.583	.197	.000	.197	.886	4.718	4.899	.181	.000	.181	.881	4.824	5.000	.176	.000	.176
	KR21 = .902, SEH_L = 4.385						KR21 = .886, SEH_L = 4.718						KR21 = .881, SEH_L = 4.824					

NOTE: Δ_1 = SEH′ − SEH; [illegible] = SEH_L − SEH; Δ_3 = SEH′ − SEH_L